Longitudinal Employer - Household Dynamics

Technical paper No. TP-2003-07

Origin-Destination Matrix and Block Characteristics Files: Prototype developed for the Bureau of Transportation Statistics

Date	:	September 2003
Prepared by	:	Marc Roemer
Contact	:	U.S. Census Bureau, LEHD Program FB 2138-3 4700 Silver Hill Rd. Suitland, MD 20233 USA

This research is a part of the U.S. Census Bureau's Longitudinal Employer-Household Dynamics Program (LEHD), which is partially supported by the National Science Foundation Grant SES-9978093 to Cornell University (Cornell Institute for Social and Economic Research), the National Institute on Aging, and the Alfred P. Sloan Foundation. The views expressed herein are attributable only to the author(s) and do not represent the views of the U.S. Census Bureau, its program sponsors or data providers. Some or all of the data used in this paper are confidential data from the LEHD Program. The U.S. Census Bureau is preparing to support external researchers' use of these data; please contact U.S. Census Bureau, LEHD Program, Demographic Surveys Division, FOB 3, Room 2138, 4700 Silver Hill Rd., Suitland, MD 20233, USA.

A Prototype

of an Origin-Destination Matrix

and Block Characteristics Files

for the Bureau of Transportation Statistics

September 30, 2003

Marc Roemer
Longitudinal Employer-Household Dynamics
Bureau of the Census
301.763.5291
marc.i.roemer@census.gov

The OD matrix contains 5 fields. The 'h' prefix denotes 'home' and 'w' denotes 'work'.

h_geocode	FIPS state, county; Census Tract 2000; home
h_block	Census Block 2000 within tract, home
w_geocode	FIPS state, county; Census Tract 2000; work
w_block	Census Block 2000 within tract, work
travelers	Number of workers making this trip

The Florida data contains 3,556,876 records.
The Illinois data contains 3,658,009 records.

The home block characteristics file contains 7 fields.

h_geocode	FIPS state, county; Census Tract 2000; home
h_block	Census Block 2000 within tract, home
workers	Number of workers living on this block
lo	Proportion of workers $0<annual earnings<=$12,000
md	Proportion of workers $12,000<annual earnings<=$35,000
hi	Proportion of workers $35,000<annual earnings
avgwages	Averages wages on this block

The Florida data contains 185,047 records.
The Illinois data contains 163,993 records.

The work block characteristics file contains 14 fields. SIC stands for Standard Industrial Classification. A "sic_" field equals '1' if an establishment in the industry operates on the block, otherwise it equals '0'.

w_geocode	FIPS state, county; Census Tract 2000; work
w_block	Census Block 2000 within tract, work
mmpay_w	Mean (average) monthly pay per worker on this block
mmpw_ds	Disclosure status of mmpay_w
sic_a	SIC division: Agriculture, Forestry, and Fishing
sic_b	SIC division: Mining
sic_c	SIC division: Construction
sic_d	SIC division: Manufacturing
sic_e	SIC division: Transportation, Communications, Electric, Gas, and Sanitary Services
sic_f	SIC division: Wholesale Trade
sic_g	SIC division: Retail Trade
sic_h	SIC division: Finance, Insurance, and Real Estate
sic_i	SIC division: Services
sic_j	SIC division: Public Administration

The Florida data contains 93,408 records.
The Illinois data contains 81,028 records.

The place of work. The place of work derives from the quarterly ES202 data supplied by Florida and Illinois to the LEHD program. The ES202 data contain characteristics of establishments by quarter. LEHD identifies establishments in the ES202 by the SEIN (the State Employer Identification Number for the firm) and SEINUNIT (a place of business within the firm). Because of inadequate address information, a small amount (about 5%) of ES202 data is excluded from geocoding and therefore from the OD matrix and block characteristics files.

LEHD uses internal Census Bureau data to maximize the accuracy of the geocodes. This prototype uses the best geography of each SEINUNIT during the 4 quarters of 2001. The best geography is a physical address coded to 'rooftop' level (most precise) in quarter 2, and the last choice is a mailing address coded to the 'tract' level (least precise) in quarter 4.

A clerical operation supplied improved addresses for 12,455 establishments in Florida and 46,605 in Illinois. A separate report gives details of this process. Some looked-up addresses that the states returned were unusable because the identifiers didn't match LEHD's files, or multiple addresses existed for an establishment (only 1 address per establishment was useable).

In Florida, the address improvement project increased the number of establishments geocoded to the rooftop or block face by 2.9 percentage points (from 63.5% to 66.4%). Employment at such precisely geocoded establishments increased by 5.5 percentage points (from 64.2% to 69.7%). In Illinois the number of establishments geocoded to the rooftop or block face increased by 10.2 percentage points (from 60.5% to 70.7%), and employment at such precisely geocoded establishments increased by 6.7 percentage points (from 70.8% to 77.5%).

The place of residence. The place of residence comes to LEHD from the Census Bureau's StARS (Statistical Administrative Records System) database. StARS encompasses several administrative data sets including federal tax forms, and it uses TIGER (Topologically Integrated Geographic Encoding Referencing) to geocode people to their homes. In 2001, the StARS database was unavailable and LEHD instead received a geocoded extract from federal tax forms. For this prototype, LEHD included any worker in the state geocoded to a Census Block (and more accurately than as a ZIP centroid or county centroid). Coarsely geocoded residences make about 10% of workers ineligible for the prototype.

The worker-employer link. The worker-employer link derives from the UI (Unemployment Insurance) wage reports supplied to LEHD by the partner states. These reports include the SEIN. For SEINs comprising more than 1 SEINUNIT, a statistical model developed at LEHD generates likely SEINUNITs within the SEIN for a worker. The model uses 4 types of information to associate workers with likely establishments: 1) the distance from the worker's home to the SEINUNIT; 2) the number of employees at the SEINUNIT; 3) the work history of the employee; and 4) the period of the SEINUNIT's existence.

This prototype linked a worker to the SEIN at which he earned the most during quarter 2 of 2001 and up to 10 implicated SEINUNITs from LEHD's statistical model. A weight equal to the proportion of times the model implicates the SEINUNIT applies to each worker-establishment link.

The OD Matrix and Confidentiality. This data set brings together the place of residence and place of work by the worker-employer link. This prototype releases a 'trip' in the OD matrix only if the block of residence had at least 5 workers living in it, and these workers traveled to a sum of at least 3 different blocks of work. This requirement makes about 10% of blocks (but only 2% of workers) ineligible for the OD matrix.

Home Block Characteristics and Confidentiality. This data set brings together the place of residence and the UI wage data. Characteristics of a home block are releaseable only if at least 5 workers live on the block. True wages for each worker is the sum of all wages earned at all employers in 2001 according to the UI wage records. A kernel density estimator (KDE) creates a statistically representative distribution of true wages on each block. The average wages and proportions of workers in the wage categories on each block derive from the statistically representative distribution.

Work Block Characteristics and Confidentiality. This data set comes entirely from the ES202. The following restrictions apply to the characteristics of work blocks LEHD can release. No more industry detail beyond SIC division is discloseable, so a set of flags indicates whether at least 1 establishment in each division operates on the block. The average monthly pay per worker on the block is the wages paid by all employers (each fuzzed by the LEHD ratio system) divided by the average monthly employment on the block. A disclosure status flag "mmpw_ds" indicates the outcome of the ratio disclosure-proofing system as follows:

mmpw_ds

-2 there are no establishments in this state, year, quarter, geography, industry category for this quarter (but there are data for other quarters)

-1 unable to compute this estimate because historical data are not available or future quarters are required

0 there is no employment in this cell, but there are establishments in the category (OK to disclose a 0)

1 OK

2 less than 3 employees (value suppressed in publications)

3 less than 3 employers (value suppressed in publications)

9 data significantly distorted

Using the data. LEHD intends for data users to aggregate the block-level data to larger geographic entities such as TAZ (Traffic Analysis Zone) or Census Tract before performing an analysis. The residence and workplace geocoding systems produce block-coding errors mostly small in distance but large in frequency. We also caution users that misreporting by employers in the ES202 causes spikes of trips to some blocks. For example, a large multi-establishment employer may report as a single establishment at a single address. Note also that the tract codes on the files are unique within the United States, but the block codes are unique only within the tract.